INDIAN JOURNAL OF APPLIED LINGUISTICS
VOL. 42, NO. 1-2, JAN-DEC 2016

Some Corpus Access Tools for Bangla Corpus

NILADRI SEKHAR DASH
Indian Statistical Institute, Kolkata

ABSTRACT

The techniques and strategies that are used to develop some Corpus Access Tools (CATs) for Bangla, as a part of Bangla Language Tool Kit (BLTK) are reported here. These tools are useful for retrieving linguistic data and relevant information from the modern Bangla written text corpus. At the initial stage only three tools are developed – Word Search Tool (WST), Collocation Search Tool (CST), and Sentence Search Tool (SST), which are combined into a single graphical user interface so that it can work quite elegantly to retrieve required data and information from the corpus. The hurdles that are encountered while trying to develop these tools are also addressed in this paper. Moreover, the problems and the solutions to these problems are also explicated here so that future researchers do not face much trouble to generate xml files to design new types of corpus access tool. One can visualize direct application of these tools in lexical search, concordance, collocation, language teaching, dictionary compilation, and language description. The strategies and techniques that have been adopted for developing these tools for the Bangla language corpus may also be utilized successfully to create xml files and similar tools for corpus processing for other Indian languages.

Keywords: corpus, Bangla, word search, collocation, sentence, part-of-speech, tagging

1. INTRODUCTION

The idea of developing Corpus Access Tools (CAT) was conceived when some Bangla corpus users asked for such